Towards Principled Data Science Assessment: The Personal Data Science Process (PdsP)

A Position Paper

Ismael Caballero[1], Laure Berti-Equille[2], and Mario Piattini[1]
[12]*Institute of Technologies and Information Systems, Univ. of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain*
[2]*Qatar Computing Research Institute, Doha, Qatar*
ismael.caballero@uclm.es, lberti@qf.org.qa, mario.piattini@uclm.es

Keywords: with the Unstoppable Advance of Big Data, the Role of Data Scientist Is Becoming More Important than Ever before, in This Position Paper, We Argue That Scientists Should Be Able to Acknowledge the Importance of Data Quality Management in Data Science and Rely on a Principled Methodology When Performing Tasks Related to Data Management, in Order to Quantify How Much a Data Scientist Is Able to Perform the Core of Data Management Activities We Propose the Personal Data Science Process (PdsP), Which Includes Five Staged Qualifications for Data Science Professionals, the Qualifications Are based on Two Dimensions: Personal Data Management Maturity (PDMM) and Personal Data Science Performance (PDSPf), the First One Is Defined According to Dgmr, a Data Management Maturity Model, Which Include Processes Related to the Areas of Data Management: Data Governance, Data Management, and Data Quality Management, the Second One, PDSPf, Is Grounded on PSP (Personal Software Process) and Cover the Personal Skills and Knowledge of Data Scientist When Participating in a Data Science Project, These Dimensions Will Allow to Developing a Measure of How Well a Data Scientist Can Contribute to the Success of the Organization in Terms of Performance and Skills Appraisal.

Abstract: Data Scientist, Maturity Model, Data Governance, Data Management, Data Quality Management.

1 INTRODUCTION

Data scientist is considered the sexiest job in the world of the 21st century (Davenport and Patil 2012). With the growth of Big Data and the increasing demand of Big Data professionals, it is becoming more important than ever to describe the required skills that any worker should have in order to perform successfully the functions related to Data Science. (Partnership 2014) classifies these skills into two groups: Hard Skills (Subject matter expertise, math and statistics knowledge and data and technical skills) and soft skills (problem solving, storytelling, collaboration, creativity, communication and curiosity) (Partnership 2014)

Due to the great impact that any decision taken on data could have for the organization, it is paramount to make available data with adequate levels of quality for the tasks at hand. This is not only a matter of reactively cleaning the data, but to make sure that data is adequately managed through its entire data lifecycle, from the sources up to the targets (Redman 2013). Data quality is often understood as *fitness for use* (Strong, Lee et al. 1997). According to this, the stakeholders using the data should be able not only to specify when the data is adequate for a given task but also to specify some specific data quality requirements for all of the stages of the lifecycle of the data science projects. This does not mean that the stakeholders should be in charge for the activities related to data management. Literature describes a set of roles in charge of specific activities related to data quality management: Chief Data Officer (Yang, Madnick et al. 2014), or data stewards (Plotkin 2013), or data governors (Seiner 2014). Data scientists should be expert in the analysis of data, but not necessarily in data management as both disciplines are core skills for data management (Pryor and Donnelly 2009). However, this does not mean that they should not know how to better perform the data management activities in order to obtain better insights from the data.

Thus, we argue that data scientists should be integrated in the data management processes since they are the most relevant source of requirements when it comes to extract the highest value and key performance indicators from the data. Therefore, we pose that data scientists' ability of being involved in any given data project management should be somehow measured in order to let organizations know how to choose the most adequate professionals for a task.

The way in which we propose the measurement of this ability covers two dimensions: on one hand, it covers the data management expertise of the data scientist; and on the other hand, his/her efficiency when performing specific data science tasks.

To define the measure of the first dimension –named as **personal data management maturity** - we ground our proposal on an existing data management maturity model: **dgmr** (Caballero, Serrano et al. 2013) which is further described in Section 2. Similarly, the measure for the second dimension – named as **personal data science performance**- is grounded on the Personal Software Process (PSP) described in (Humphrey 2000) (1997) as "*a set of methods, forms, and scripts that show software engineers how to plan, measure, and manage their work*"

With these two dimensions, we propose the *Personal Data Scientist Process* as a structured set of process descriptions, measurements, and methods that can help data scientists to improve their personal performance and their ability to act and decide on the various steps of the lifecycle of data used to conduct the various analyses.

To the best of our knowledge, no one has ever proposed any principled methodology for data scientist (self-) appraisal. The main rationale for this proposal is to develop a universal recognition of the skills and capabilities of professional working on Data Science. In this sense, organizations can select the most valuable professionals for their projects, and data scientists can self-assess themselves against a common reference framework.

The remainder of the paper is structured as follows: Section 2 introduces the most important concepts underlying our position paper. Section 3 describes the PdsP. Section 4 introduces an illustrative example to describe the framework. Finally, Section 5 provides conclusions and future work.

2 STATE OF THE ART

In this section, we introduce the most relevant concepts to better ground the basis of our proposal.

2.1 Required Skills for Data Scientists

Data scientists usually have a strong educational background in Mathematics, Statistics, Computer Science or Engineering. They can acutely understand the business problems and needs of the industry they are working in and fluently translate their technical findings to a non-technical team, such as the Marketing or Sales departments. Along with strong technical skills in Analytics (mastering R or SAS) data scientists should have skills in Computer Science for big data management and experience with Hadoop platform, Pig or Hive and also be able to write complex SQL queries. Their goal is to arm the business and decision makers with quantified insights for their decision-making process and technical skills to tame, clean, and analyse the data appropriately.

2.2 Dgmr Framework

This section briefly introduces dgmr, which is a framework containing three main elements:

- A process reference model, describing the processes related to data management (DM), data quality management (DQM) and data governance (DG). These processes are described as ISO 12207 does. See Table 1.
- A maturity model, in which the processes previously described, has been arranged in five levels, according to what organizations should perform in order to maintain the highest levels of quality and availability for data. See Figure 1.
- An assessment methodology, which enables the assessment of the level of organizational data management maturity.

3 PDSP

- The PdsP describes the concepts and processes that any data scientist should learn and follow to get a better job when analysing data. In this context, "a better job" means not only getting more reliable results but also more repeatable results in a more productive way.

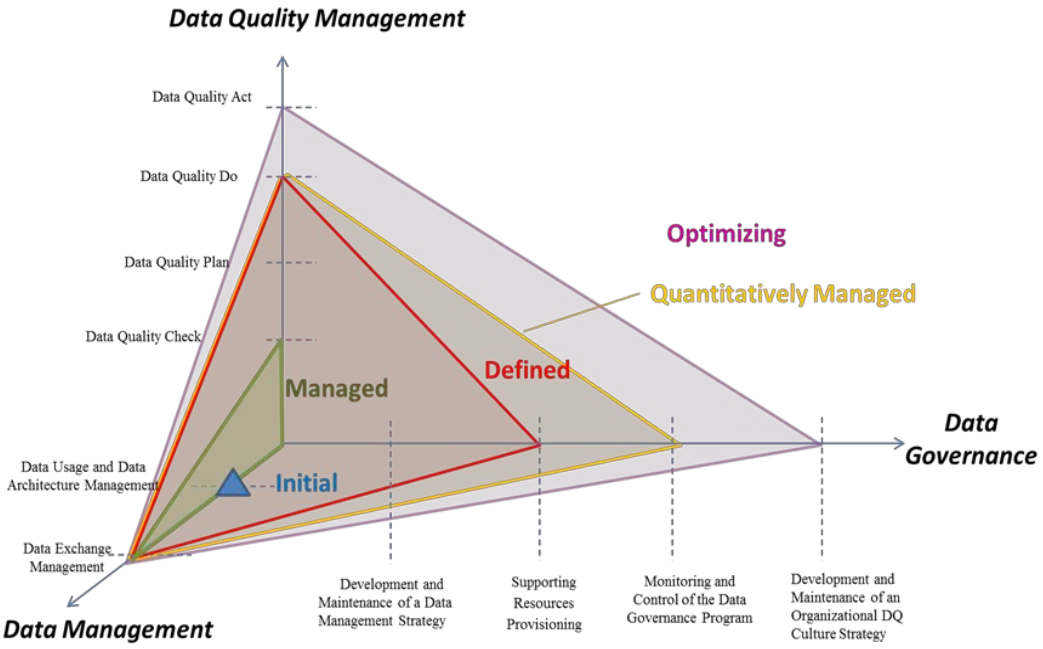

Figure 1: Data Management Maturity Model associated to dgmr.

Table 1: dgmr.

Disci	Areas
DM	Data Usage and Data Architecture Management
	Data Exchange Management
DQM	Data Quality Plan
	Data Quality Do
	Data Quality Check
	Data Quality Act
DG	Development and Maintenance of a Data Management Strategy
	Development and Maintenance of an Organizational DQ Culture Strategy
	Supporting Resources Provisioning
	Monitoring and Control of the Data Governance Program

The design of PdsP is based on analogous principles as PSP (Humphrey 2000):

1. Every data scientist is different; to be most effective, data scientists should be able to plan their work and they should base their plan on their own personal data.
2. In order to improve their performance, data scientists should follow well-defined and measured processes.
3. High quality data analysis must be achieved by highly motivated and responsible data scientist.
4. Data scientists must know the context of the data used in the data science project as well as the whole data lifecycle and the data science project lifecycle.
5. It is usually cheaper to find and defects in data analysis earlier than later.
6. It is more efficient to prevent defects than to find and fix them.
7. The right way is always the fastest and cheapest way to do a job

3.1 Dimensions of PdsP

How well data scientists perform in these principles is measured by using two dimensions, namely:

- **Personal data management maturity** (PDMM). It measures the extent to which a data scientist can understand the data management foundations. This data management foundations include data management (referring to data management itself), data quality and data governance. This extent is aligned to dgmr. Coherently, we define five possible values to represent the skills and knowledge of the data scientists: initial, managed, defined, quantitatively managed and optimizing. These values reflect how well a data scientist can address his/her tasks within the data lifecycle. This is important from the business and IT point of view because enable data scientist to better contextualize the data he/she is using and to understand the business value of the data for the task. See Figure 1.
- **Personal data science performance (PDSPf)**. It represents the extent to which a data scientist is disciplined when they conduct the various data analysis. In this sense this dimensions measures how much the data scientist follow the best-practices in order to produce high quality results in a predictable way and within schedule and budget. In order to quantify this dimension, and analogously as PSP does, we define four sets of processes containing the best practices for data analysis. The sets are listed in growing order. Each immediately higher level includes the best practices of the previous one and some other new ones which implies a higher personal maturity for the data scientist when performing data analysis. See Table 2 for the description of the sets. It is important to realize that the performance of data scientist is measured in terms of time, size and quality measures.

Therefore, depending on the ability of the data scientist of meeting the best practices of a specific set, this will be his/her measure for PdsPf.

3.2 How to Determine the PdsP Level

According to the two dimensions previously explained, we identify five qualifications for professionals – PdsP level, which can be inferred by plotting these two dimensions across (please, see Figure 2).

Table 2: Best practices for each PdsPf levels.

Set	Best practices addressed
PdsPf0	Meeting requirements of the task Performing basic measurement about the execution of the task Using Coding standards Process improvement proposal Size measurement
PdsPf1	Size Estimating Test Report Task Planning Schedule planning
PdsPf2	Data reviews Design review Design and use of templates
PdsPf3	Cyclic development

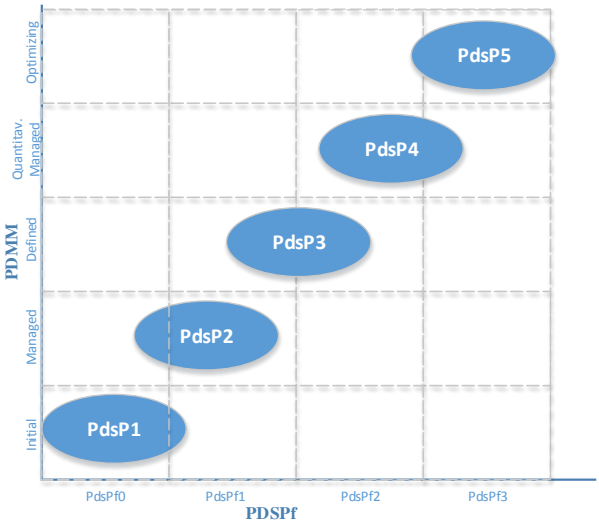

Figure 2: Values for PdsP (Caballero, Serrano et al. 2013).

The five ideal levels of qualification we propose in the PdsP framework are:

- **PdsP1** corresponds to a junior profile. A data scientist of this type will be able to perform easy tasks and he/she can understand the role of the data in the data analysis and how to interact with the data architecture to recover and store data, but maybe he/she is not still ready enough to become involved in any data management task.
- **PdsP2**, in which a data scientist is able to perform basic data analysis and can slightly

contribute with some requirements related to data management. He/she can understand the importance of data quality for data used in the analysis, but he/she cannot describe how to achieve it.

- **PdsP3** is an intermediate profile of somebody who can contribute efficiently to tasks related to data management and data quality management and is able to perform complex analysis on the data. He/she fully understand the meaning of data quality and how to achieve it by following data management strategy that he/she will take into account during the design of the data analysis tasks.
- **PdsP4** corresponding to data scientist profile who can perform very complex data analysis and provides requirements for optimizing the analysis and monitoring data management, data quality management, and data governance.
- **PdsP5** corresponds to a senior profile. The Data Scientist is not only able to lead data science projects, but he/she can also provide end-to-end data governance requirements.

However, it is possible to find profiles of data scientist with a very low PDMM and a very high PDSPf. Professionals in this situation are supposed to know very well the data in the organization although they are not able to face up with any managerial task.

4 CONCLUSIONS

This paper has presented the Personal Data Scientist Process, which can help Data Scientist to get more mature. In fact, the benefits that any Data Scientist could expect from the application of the framework are:

- It can help the data scientists in better developing high quality data products from the corresponding analysis
- It can better guide data scientists for personal improvement
- It gives data scientists the command over the work they are performing
- It gives not only the necessary confidence on herself to perform a better job but also to improve the ways of doing added-value activities.

As future work, we will work on the following concerns: define and adapt the scripts provided in PSP to PDSPf, define and test equivalent metrics for size and quality in data science project. Also we want to conduct some case studies to better delimit the scope of each PdsP level and to introduce the framework to different organizations in order to validate it.

ACKNOWLEDGEMENTS

This work has been partially funded by the VILMA project (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-048-P) and by the GEODAS-BC project (Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional FEDER, TIN2012-37493-C03-01)

REFERENCES

Caballero, I., S. Serrano and M. Piattini (2013). Technical Report: dgmr: A Data Management Maturity Model, University of Castilla-La Mancha.

Davenport, T. H. and D. Patil (2012). "Data Scientist." Harvard Business Review **90**: 70-76.

Humphrey, W. S. (2000). Introduction to the Personal Software Process (PSP). Pittsburg (Pennsylvania, USA), Addison Wesley.

Partnership, T. (2014). Big Data Analytics: Assessment of Demand for Labout and Skills 2013-2020. T. Partnership. London (UK), SAS UK & Ireland.

Plotkin, D. (2013). Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance, Newnes.

Pryor, G. and M. Donnelly (2009). "Skilling up to do data: whose role, whose responsibility, whose career?" International Journal of Digital Curation **4**(2): 158-170.

Redman, T. (2013). Data Quality Management Past, Present, and Future: Towards a Management System for Data. Handbook of Data Quality. S. Sadiq, Springer**:** 15-40.

Seiner, R. S. (2014). "Non-invasive data governance.".

Strong, D. M., Y. W. Lee and R. Y. Wang (1997). "Data quality in context." Communications of the ACM **40**(5): 103-110.

Yang, L., S. Madnick, R. Wang, F. Wang and Z. Hongyun (2014). "A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data." MIS Quarterly Executive **13**(1): 1-13.